TECOLOTE RESEARCH, INC.

Bridging Engineering and Economics

Since 1973

Fit, Rather Than Assume, a CER Error Distribution

18 December 2013

Dr. Shu-Ping Hu

■ Los Angeles ■ Washington, D.C. ■ Boston ■ Chantilly ■ Huntsville ■ Dayton ■ Santa Barbara

■ Albuquerque ■ Colorado Springs ■ Goddard Space Flight Center ■ Johnson Space Center ■ Ogden ■ Patuxent River ■ Washington Navy Yard

■ Ft. Meade ■ Ft. Monmouth ■ Dahlgren ■ Quantico ■ Cleveland ■ Montgomery ■ Silver Spring ■ San Diego ■ Tampa ■ Tacoma

■ Aberdeen ■ Oklahoma City ■ Eglin AFB ■ San Antonio ■ New Orleans ■ Denver ■ Vandenberg AFB

Objectives

- **Recommend using an objective approach, not an assumption, to model CER error distributions**
  - A hypothesized distribution (e.g., normal, log-normal, triangular, etc.) may not be appropriate to model the errors of a cost estimating relationship (CER) for cost uncertainty analysis
- **Develop easy-to-follow guidance for analysts to derive distribution fitting results for cost uncertainty analysis**
  - The fitted distribution should be adjusted properly to build prediction intervals for cost uncertainty analysis

Our goal is to derive CER error distributions from real data rather than from assumptions

Outline

- Objectives
- Common Questions for Fitting CER Errors
- Prediction Interval (PI) Analysis
- Adjustment Factors for Uncertainty Analysis
- Easy-to-Follow Implementation Steps
- Concerns about Analyzing Different CER Errors Together
- Analyzing Errors for USCM9 Subsystem-Level CERs
- Conclusions and Recommendations

Common Questions for Fitting CER Errors (1/3)

- **What should we analyze for (ordinary least squares) OLS CERs?**
  - residuals $(y_i - \hat{y}_i)$
  - standardized residuals $((y_i - \hat{y}_i)/se(y_i - \hat{y}_i))$

y_i : Actual Observation
$\hat{y}_i$: CER Predicted Value

$i = 1,\ldots, n$
n = sample size

- **What should we analyze for MUPE and ZMPE CERs?**
  - ratios of actual to predicted $(y_i/\hat{y}_i)$
  - percentage errors $((y_i - \hat{y}_i)/\hat{y}_i)$
- **Findings:**
  - Just like residual vs. standardized residual plots, the histograms of residuals and standardized residuals look very similar. It is adequate to fit residuals to find the error distribution for additive CERs.
  - Percentage errors are centered on zero; hence, they cannot be fitted by a log-normal distribution unless a location parameter is used

Analyze (1) residuals $(y_i - \hat{y}_i)$ for additive error models and (2) ratios of $y_i/\hat{y}_i$ for MUPE and ZMPE CERs

Common Questions for Fitting CER Errors (2/3)

- **What should we analyze for log-error CERs, $y_i/\hat{y}_i$ in unit or log space?**
- **Two methods are commonly used to fit a log-normal distribution**
  - **Maximum-Likelihood Estimation (MLE) solution** for μ and σ in log space

$$\hat{\mu} = \frac{\sum_{i=1}^{n}\left(\ln(y_i) - \ln(\hat{y}_i)\right)}{n}$$

$$\hat{\sigma} = \frac{\sum_{i=1}^{n}\left(\ln(y_i/\hat{y}_i) - 0\right)^2}{n} = \frac{\sum_{i=1}^{n}\left(\ln(y_i) - \ln(\hat{y}_i)\right)^2}{n}$$

$\hat{\mu}$ and $\hat{\sigma}$ are evaluated in log space; $\hat{\mu}$ should be zero for log-linear CER. CB uses (n-1) in the denominator to estimate σ; @Risk uses the sample size n. It should be (n-p) to account for DF.

  - **"Least Square" solution** for μ and σ in unit space

$$Minimizing \sum_{i=1}^{n}\left((y_i/\hat{y}_i) - Loginv((0.5*\text{ObsFreq} + \text{NumObsBelow})/n, \mu, \sigma)\right)^2$$

where ObsFreq = the number of sample points equal to y_i, inclusive

NumObsBelow = the number of observations below the value of y_i

  - **MLE and Unit-space Least Square solutions are different**

Fit ratios of $y_i/\hat{y}_i$ in log space for log space OLS (LOLS) CERs for consistency

Common Questions for Fitting CER Errors (3/3)

- **Should we apply any adjustments to the distribution fitting tool results for uncertainty analysis?**
- **Findings:**
  - We should apply adjustments when fitting distributions to CER errors, as well as sample data. Otherwise, the range of the PI will be smaller than it should be

Adjustments should be applied when using distribution fitting tool results for uncertainty analysis

Use prediction interval (PI) concept to derive adjustments for CER uncertainty analysis when using a distribution fitting tool

Uncertainty Analysis
PI for OLS: Y = a + bX + ε (ε ~N(0, Iσ²))

- **A (1-α)100% PI for OLS is given below when X = x_o** (an estimating point):

$$PI = \hat{y}_0 \pm (t_{\alpha/2,n-2}) SE \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_{xx}}} = \hat{y}_0 \pm (t_{\alpha/2,n-2}) * (Adj.SE)$$

Adjusted SE for OLS

The PI formula can be extended to include multiple driver variables

  - x_o is the value of the predictor variable used in calculating the estimate
  - $\hat{y}_0$ is the estimated value from the CER when X = x_o
  - **SE** is CER's standard error of estimate; "n-2" is degrees of freedom (DF)
  - "Adj. SE" is the adjusted standard error for PI
  - $t_{(\alpha/2,\ n-2)}$ is the upper $\alpha/2$ cut-off point for a t distribution with "**n-2**" DF
  - $\bar{x} = (\sum_{i=1}^{n} x_i)/n$ and $SS_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2$

- **Use Student's t distribution to model CER uncertainty: enter the "Adj. SE" in the scale field and specify DF in the degrees of freedom field**
- **If the data set is unavailable, we can use a heuristic approach to approximate the "Adj. SE" measure**

Uncertainty Analysis

PI for WLS: $Y = \alpha + \beta X + \varepsilon = f(X) + \varepsilon$ ($\varepsilon \sim N(0, V\sigma^2)$)

- **A (1-α)100% PI for WLS when X = x_o** (an estimating point):

The PI formula can be extended to include multiple driver variables

Adj. SE for WLS

$$PI = f(x_0) \pm t_{(\alpha/2,\, n-2)} * SE * \sqrt{\frac{1}{w_o} + \frac{1}{\sum w_i} + \frac{(x_0 - \bar{x}_w)^2}{SS_{wxx}}} = f(x_0) \pm t_{(\alpha/2,\, n-2)} * (Adj.\ SE)$$

  - $f(x_o)$, i.e., $\hat{y}_0$, is the estimated value from the CER when X = x_o
  - w_o is the weighting factor for y when x = x_o ($w_o = (1/f(x_o))^2$ for MUPE)
  - w_i is the weighting factor for the i^{th} data point ($w_i = 1/(f(x_i))^2$ for MUPE)
  - "Adj. SE" is the adjusted standard error for PI
  - $t_{(\alpha/2,\, n-2)}$ is the upper $\alpha/2$ cut-off point for a t distribution with "**n-2**" DF
  - $\bar{x}_w = \sum_{i=1}^{n} w_i (x_i) \Big/ \sum_{i=1}^{n} w_i$ and $SS_{wxx} = \sum_{i=1}^{n} w_i (x_i - \bar{x}_w)^2$

- **Use Student's t distribution to model CER uncertainty: enter the "Adj. SE" in the scale field and specify DF in the degrees of freedom field**
- **If the data set is unavailable, we can use a heuristic approach to approximate the "Adj. SE" measure**

Uncertainty Analysis
PI for MUPE Factor CER: Y = βX * ε (ε ~N(0,Iσ²)) (1/2)

- **A (1-α)100% PI for MUPE Factor CER when X = x_o** (an estimating point):

$$PI = \hat{y}_0 \pm t_{(\alpha/2,\,n-1)} * SE * \sqrt{\frac{1}{w_0} + \frac{(x_0)^2}{\sum_{i=1}^{n} w_i x_i^2}}$$

Adjusted SE for a weighted factor CER

Adjusted SE for a MUPE/ZMPE factor CER

$$= \hat{y}_0 \pm t_{(\alpha/2,\,n-1)} * SE * \sqrt{(1+\frac{1}{n}) b^2 x_0^2} = \hat{y}_0 \left(1 \pm t_{(\alpha/2,\,n-1)} * SE * \sqrt{1+\frac{1}{n}} \right) = \hat{y}_0 (1 \pm t_{(\alpha/2,\,n-1)} * Adj.\,SE)$$

  - $\hat{y}_0$ (=bx_o) is the estimated value from the CER when X = x_o
  - w_o is the weighting factor for y when x = x_o (w_o = 1/(bx_o)^2 for MUPE)
  - w_i is the weighting factor for the ith data point (w_i =1/(bx_i)^2 for MUPE)
  - "Adj. SE" is the adjusted standard error for PI
  - $t_{(\alpha/2,\,n-1)}$ is the upper α/2 cut-off point for a t distribution with "**n-1**" DF
- **Use Student's t distribution to model CER uncertainty: enter the "Adj. SE" in the scale field and specify DF in the degrees of freedom field**
- **We do not need the actual data set to a build PI for MUPE and ZMPE factor CERs since the adjustment is a constant factor**

Uncertainty Analysis
PI for MUPE Factor CER: Y = βX * ε (ε ~N(0,Iσ²)) (2/2)

- **A (1-α)100% PI for MUPE Factor CER when X = x_0** (an estimating point):

$$PI = \hat{y}_0\left(1 \pm t_{(\alpha/2,\,n-1)} * SE * \sqrt{1+\frac{1}{n}}\right) = bx_0\left(1 \pm t_{(\alpha/2,\,n-1)} * \frac{S_z}{\bar{z}} * \sqrt{1+\frac{1}{n}}\right) = bx_0\left(1 \pm t_{(\alpha/2,\,n-1)} * (Adj.\,SE)\right)$$

Adjusted SE for a MUPE/ZMPE factor CER

  - *b* is the estimated factor for MUPE/ZMPE CER; z = y/x
  - SE = $S_z/\bar{z}$ = CV(Z) where z = y/x and S_z is the standard deviation of Z

$$(SE)^2 = \frac{1}{n-1}\sum_{i=1}^{n} w_i (y_i - bx_i)^2 = \frac{1}{n-1}\sum_{i=1}^{n} \frac{1}{(bx_i)^2}(y_i^2 - 2bx_i y_i + (bx_i)^2)$$

$$= \frac{1}{n-1}\left(\frac{1}{b^2}\sum_{i=1}^{n}\left(\frac{y_i}{x_i}\right)^2 - \frac{2}{b}\sum_{i=1}^{n}\frac{y_i}{x_i} + n\right) = \frac{1}{n-1}\left(\frac{1}{b^2}\sum_{i=1}^{n}\left(\frac{y_i}{x_i}\right)^2 - \frac{2}{b}(nb) + n\right)$$

$$= \frac{1}{b^2(n-1)}\left(\sum_{i=1}^{n}\left(\frac{y_i}{x_i}\right)^2 - nb^2\right) = \frac{1}{b^2}\frac{1}{(n-1)}\left(\sum_{i=1}^{n} z_i^2 - n\bar{z}^2\right)$$

$$= \frac{(S_z)^2}{\bar{z}^2} = (CV(Z))^2$$

$w_i = 1/(bx_i)^2$ for MUPE/ZMPE

$$b = \left(\sum_{i=1}^{n}\frac{y_i}{x_i}\right)/n = \bar{z}$$

- **Note: the PI for MUPE (and ZMPE) factor CER can be expressed by a simple closed form formula**

Uncertainty Analysis
PI for LOLS: $Y = \alpha * X^{\beta} * \varepsilon$ $(\varepsilon \sim LN(0, I\sigma^2))$

- **A (1-α)100% PI for LOLS is given below when X = x_o** (an estimating point):

$$PI = Exp\left(\hat{y}_{\log} \pm (t_{\alpha/2,\, n-2}) * SE * \sqrt{1 + \frac{1}{n} + \frac{(\ln(x_0) - \overline{\ln(x)})^2}{\sum_{i=1}^{n} (\ln(x_i) - \overline{\ln(x)})^2}} \right)$$

$$= Exp\left(\hat{y}_{\log} \pm (t_{\alpha/2,\, n-2}) * (Adj.SE) \right)$$

Adjusted SE for LOLS

The PI formula can be extended to include multiple driver variables

  - x_o is the value of the predictor variable used in calculating the estimate
  - $\hat{y}_{log}$ is the estimated value in log space when X = x_o
  - **SE** is CER's standard error of estimate in **log** space
  - $\overline{\ln(x)}$ is the average of all the values of x_i's evaluated in log space
- **Use Log-t distribution to model CER uncertainty: enter "Adj. SE" in the scale field and specify DF in the degrees of freedom field**
- **If Log-t distribution is not available, use Student's t distribution in log space, but make sure to bring the results back to unit space**

Tip: Use Log-T distribution to construct PI for LOLS CERs

Uncertainty Analysis
PI for Univariate Analysis

- **Given a random sample $\{y_1, y_2, \ldots, y_n\}$ from a normal distribution, a $(1-\alpha)100\%$ PI for a future observation is given by**

$$PI = \bar{y} \pm t_{(\alpha/2,\, n-1)} * S_Y * \sqrt{1+\frac{1}{n}} = \bar{y}\left(1 \pm t_{(\alpha/2,\, n-1)} * \frac{S_Y}{\bar{y}} * \sqrt{1+\frac{1}{n}}\right)$$

$$= \bar{y}\left(1 \pm t_{(\alpha/2,\, n-1)} * (Adj.\, SE)\right)$$

Adj. SE for Univariate

  - $\bar{y} = (\sum_{i=1}^{n} y_i)/n$ is the sample mean
  - $S_Y = (\sum_{i=1}^{n} (y_i - \bar{y})^2)/(n-1)$ is the sample standard deviation
  - "Adj. SE" is the adjusted standard error for PI
  - $t_{(\alpha/2,\, n-1)}$ is the upper $\alpha/2$ cut-off point for a t distribution with "n-1" DF
- **Use Student's t distribution to model the uncertainty: enter the "Adj. SE" in the scale field and specify DF in the degrees of freedom field**
- **The PI for univariate analysis is the same as the PI for the MUPE/ZMPE factor CER**

Two factors can be easily identified using the PI formula:
1) location factor (from SE to Adj.SE)
2) $t_{(\alpha/2,\ DF)}$ (from normal to t distribution)

There is a third one: regression factor

Adjustment Factors for Uncertainty Analysis (1/4)

- **A distribution fitting tool does not know**
  - whether the data set is an entire population or a random sample
  - how many coefficients are estimated by the CER (when modeling the CER errors)
- **Regression Adjustment Factor is given by**

$$\text{Regression Adjustment Factor} = \sqrt{\frac{n}{df}}$$

- **Use Regression Adjustment Factor to account for**
  - the difference between sample and population
  - the appropriate degrees of freedom if certain parameters are estimated by the sample
  - Note: "*df*" stands for the degrees of freedom of the CER

Adjustment Factors for Uncertainty Analysis (2/4)

- **Use Location Factor to account for the distance of the estimating point (i.e., x_o) from the center of the database**
  - In a simple linear model, the location adjustment factor is given by

$$\text{Location Factor} = \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_{xx}}} = \sqrt{1 + \frac{1}{n} + \frac{((x_0 - \bar{x})/S_x)^2}{n}}$$

$$S_x = \sqrt{SS_{xx}/n} = \sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2/n}$$

    where x_0 is the value of the predictor variable used in calculating the estimate and S_x is the uncorrected sample standard deviation
  - PI gets larger when the estimating point moves farther away from the center of the database
  - If the data set is unavailable, we can use a heuristic approach to approximate the "Adj. SE" measure:

$$\textit{Heuristic Assessment}: \quad \frac{\textit{Distance}}{\textit{Driver Stdev}} = \begin{cases} 0.25 & \textit{Very Similar} \\ 0.75 & \textit{Similar} \\ 1.50 & \textit{Somewhat Different} \\ 3.00 & \textit{Very Different} \end{cases}$$

Distance = $(x_0 - \bar{x})$
Driver Stdev = S_x

Adjustment Factors for Uncertainty Analysis (3/4)

- **Use DF Factor to account for small samples**

$$\text{DF factor} = \sqrt{\frac{df}{df - 2}}$$

Note: "*df*" stands for the degrees of freedom of the CER, which is the DF of Student's t (or Log-t) distribution

  - The DF adjustment factor accounts for the broader tails of Student's t (or Log-t) distribution for small samples. For example, we should multiply the *Adj. SE* by the DF factor if we use normal instead of t distribution for uncertainty analysis.
  - The DF factor is the standard deviation of a Student's t distribution with a scale parameter one and "df" degrees of freedom
  - Do not apply the DF adjustment factor if a Student's t or a Log-t distribution is chosen to model the CER errors

- **Consider applying DF, Regression, and Location Factors when using a distribution fitting tool for cost uncertainty analysis. Otherwise, the range of the PI will be smaller than it should be.**

Adjustment Factors for Uncertainty Analysis (4/4)

Location Factor by Model Type

Model	Location Factor = (Adj. SE) /SE (for one predictor variable)
Additive	Linear: $\sqrt{1+\frac{1}{n}+\frac{((x_0-\bar{x})/S_x)^2}{n}}$ Factor: $\sqrt{1+\frac{x_0^2}{\sum x_i^2}}$
Log-Linear	$\sqrt{1+\frac{1}{n}+\frac{(\ln(x_0)-\overline{\ln(x)})^2}{\sum_{i=1}^{n}(\ln(x_i)-\overline{\ln(x)})^2}}$
MUPE (Linear)	$\sqrt{1+\frac{1}{\hat{y}_0^2\sum w_i}+\frac{(x_0-\bar{x}_w)^2}{\hat{y}_0^2(SS_{wxx})}}$
MUPE (Factor) Univariate	$\sqrt{1+\frac{1}{n}}$
Heuristic	$\sqrt{1+\frac{1}{n}+\frac{(Distance/Driver\ Stdev)^2}{n}}$

Heuristic Assessment:

$$\frac{Distance}{Driver\ Stdev}=\begin{cases}0.25 & Very\ Similar\\ 0.75 & Similar\\ 1.50 & Somewhat\ Different\\ 3.00 & Very\ Different\end{cases}$$

- x_0 is the value of the independent variable used in calculating the estimate and $\hat{y}_0$ is the estimated value from the CER when X = x_0
- Distance = ($x_0 - \bar{x}$); Driver Stdev = $S_x = \sqrt{SS_{xx}/n} = \sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2/n}$ (uncorrected stdev)

Easy-to-Follow Implementation Steps (1/3)

- **Summary Table of Adjustments:**

Model Type	Adjustments
Additive	$(y_i - \hat{y}_i)$ * (Net Factor)
Log-Error	$(\ln(y_i) - \ln(\hat{y}_i))$ * (Net Factor)
MUPE/ZMPE	$(y_i / \hat{y}_i)$ * (Net Factor) – Shift
Univariate	$(y_i / \bar{y})$ * (Net Factor) – Shift

- **Net Factor (NF) = (DF Factor) * (Regression Factor) * (Location Factor)**
  - Do not apply the DF factor to compute NF if (1) deg of freedom > 50 or (2) a Student's t or a Log-t distribution is chosen to model the CER error distribution
- **Shift = NF – 1**
  - Shift is applied to (1) MUPE and ZMPE CERs to ensure the fitted distribution is centered on **1** and (2) univariate analysis to preserve the sample mean

Tip: Make appropriate adjustments before using a distribution fitting tool

Easy-to-Follow Implementation Steps (2/3)

Model Type	Adjustments
Additive	$(y_i - \hat{y}_i)$ * (Net Factor)
Log-Error	$(\ln(y_i) - \ln(\hat{y}_i))$ * (Net Factor)
MUPE/ZMPE	$(y_i / \hat{y}_i)$ * (Net Factor) – Shift
Univariate	$(y_i / \bar{y})$ * (Net Factor) – Shift

- **For consistency, we should know how the CERs/PERs were derived**
  - Fit residuals for additive models
  - Fit residuals in log space for log-error models; e.g., log-linear CERs
  - Fit percentage errors in ratios of y_i to $\hat{y}_i$ for MUPE and ZMPE CERs
- **Deduce the fitting hypothesis if the method is unknown:**
  - $\Sigma(y_i - \hat{y}_i) = 0$ ➔ OLS
  - $\Sigma(\ln(y_i) - \ln(\hat{y}_i)) = 0$ ➔ LOLS
  - $(\Sigma(y_i - \hat{y}_i)/\hat{y}_i) / n = 1$ ➔ MUPE or ZMPE (or LOLS with PING Factor or Smearing Estimate)

Easy-to-Follow Implementation Steps (3/3)

Model Type	Adjustments
Additive	$(y_i - \hat{y}_i)$ * (Net Factor)
Log-Error	$(\ln(y_i) - \ln(\hat{y}_i))$ * (Net Factor)
MUPE/ZMPE	$(y_i / \hat{y}_i)$ * (Net Factor) – Shift
Univariate	$(y_i / \bar{y})$ * (Net Factor) – Shift

- **Suggest using an additional cell for the error distribution besides PE**
  - Make sure the error term is applied to the PE appropriately
- **Be careful when using one cell for both the PE and error term**
  - Mean = PE, SD (for Student's t) = σ_u (from curve-fitting tool) * PE
  - Mean = PE, Mode (for Triangular) = 3*PE – Min* PE – Max* PE
  - Mean = PE, σ in log space (for Log-normal) = $\sqrt{\ln(1+\sigma_u^2)}$
  - Median = PE, scale parameter (for Log-t) = σ (in log space) for log-error model

For MUPE and ZMPE CERs

A MUPE CER Example: Cost = a + b*Wt (1/2)

Model Type	Adjustments
MUPE/ZMPE	$(y_i / \hat{y}_i)$ * (Net Factor) – Shift

A MUPE CER: Cost = 220.0895 + 3.8112 * Weight (SE = 28.13%, N = 49)

- **Given:** x_0 = 300 lbs, $\hat{y}_0$ = 1,363.45, SS_{wxx} = 1.072, $\bar{x}_w$ = 469.475, and Σw_i = 8.2795*10^-6
- **DF, Regression, and Location Factors are given by**
  - DF Factor = sqrt(47/45) = 1.022
  - Regression Factor = sqrt(49/47) = 1.0211
  - Location Factor $= \sqrt{1 + \frac{1}{\hat{y}_0^2 \sum w_i} + \frac{(x_0 - \bar{x}_w)^2}{\hat{y}_0^2 (SS_{wxx})}} = \sqrt{1 + \frac{10\text{^}6}{(1346.45)^2 (8.27953)} + \frac{(300 - 469.4747)^2}{(1346.45)^2 (1.07202)}} = 1.03893$

The heuristic location factor is **1.032** using 1.5 as the distance ratio to address the similarity between the estimating system and the CER database.

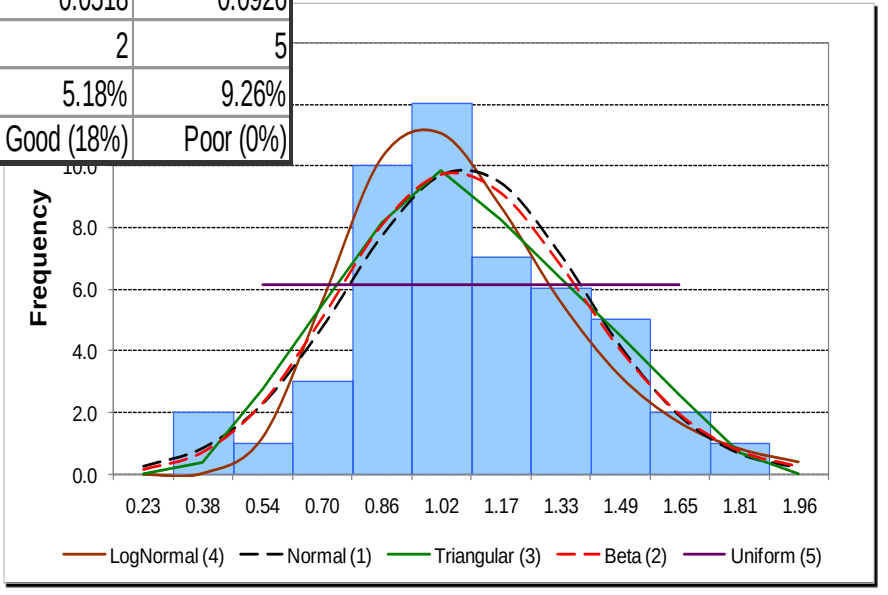
- **Net Factor = 1.0211*1.022*1.038933 = 1.084125**
- **Shift = Net Factor – 1 = 0.084125**
- **Fit: $(y_i / \hat{y}_i)$ * (Net Factor) – Shift = $(y_i / \hat{y}_i)$ * (1.084125) – 0.084125**

A MUPE CER Example: Cost = a + b*Wt (2/2)

- **Results derived by Distribution Finder for the "adjusted % errors":**

	Sample	LogNormal	Normal	Triangular	Beta	Uniform
Mean	1.0000	1.0035	1.0000	1.0000	1.0002	1.0000
StdDev	0.3125	0.3009	0.3093	0.3049	0.3078	0.2957
CV	0.3125	0.2998	0.3093	0.3049	0.3078	0.2957
Min	0.2255			0.3005	-0.6144	0.4878
Mode		0.8819	1.0000	0.9130	0.9734	
Max	1.8066			1.7866	3.8257	1.5122
Alpha					17.1439	
Beta					30.0000	
Data Count	49	% < 0 =	0.06%	None	0.01%	None
Standard Error of Estimate		0.0679	0.0504	0.0584	0.0518	0.0926
Rank		4	1	3	2	5
SEE / Fit Mean		6.76%	5.04%	5.84%	5.18%	9.26%
Chi^2 Fit test 9 Bins, Sig 0.05		Good (43%)	Good (32%)	Good (31%)	Good (18%)	Poor (0%)

Normal distribution is ranked #1 with an estimated standard deviation of 0.3093, which is almost the same as the number reported in the regression PI output.



Concerns about Analyzing Different CER Errors Together (1/2)

- **The CER errors from different CERs may not be identically distributed**
  - For example, the distribution of errors from the Structure CER may not be the same as the distribution of errors from the Electrical Power Subsystem (EPS) CER
- **The CER errors associated with different subsystems might not be independently distributed either**
  - We should examine whether or not these CER errors are correlated before pooling them together
- **This approach may not be feasible when fitting a distribution with three or more parameters**
  - **Beta** distribution: the alpha, beta, Low, and High parameters for the error distributions may not be the same across different CERs, even if all the normalized CER errors have the same mean and same variance
  - **Log-normal** distribution: we cannot define a global location parameter (in a meaningful way) for a shifted log-normal distribution when analyzing the "normalized" errors for several different CERs all together

Concerns about Analyzing Different CER Errors Together (2/2)

- **If $X \sim LN(\mu, \sigma^2)$, i.e., $LN(\mu, \sigma^2, 0)$, then $Y = aX + b \sim LN(\mu + \ln(a), \sigma^2, b)$**
  - $LN(\mu, \sigma^2, b)$ denotes a shifted log-normal distribution with a mean of μ, variance σ^2 (both in log space), and a location parameter b (in unit space)
- **Consider k different MUPE (or ZMPE) CERs:**

  $y_i = f_i \varepsilon_i$ where $E(\varepsilon_i) = 1$, $Stdev(\varepsilon_i) = \sigma_{iu}$, & $\varepsilon_i \sim LN(\mu_i, \sigma_i^2)$ for i = 1, ..., k
  - $\mu_i = -\sigma_i^2/2$ and $\sigma_i = \sqrt{\ln(1+\sigma_{iu}^2)}$
  - $(y_i - \hat{y}_i)/\hat{y}_i = (\hat{\varepsilon}_i - 1) \sim LN(\mu_i, \sigma_i^2, -1) = LN(-\sigma_i^2/2, \sigma_i^2, -1)$
  - $e_i = ((y_i - \hat{y}_i)/\hat{y}_i)/\sigma_{iu} = (\hat{\varepsilon}_i - 1)/\sigma_{iu} \sim LN(-\sigma_i^2/2 - \ln(\sigma_{iu}), \sigma_i^2, -1/\sigma_{iu})$

  σ_i is in log space
  σ_{iu} is in unit space
- **Properties of these normalized percentage errors (e_i's):**
  - $E(e_i) = 0$ and $Stdev(e_i) = 1$ for k different CERs (i = 1,…, k)
  - e_i's do **not** have the same mean and variance in **log** space; their location parameters are also different
- **e_i's should not be analyzed together using a distribution fitting tool**
  - The analysis results will be misleading and inaccurate if we combine these e_i's (from different CERs) and analyze them all together

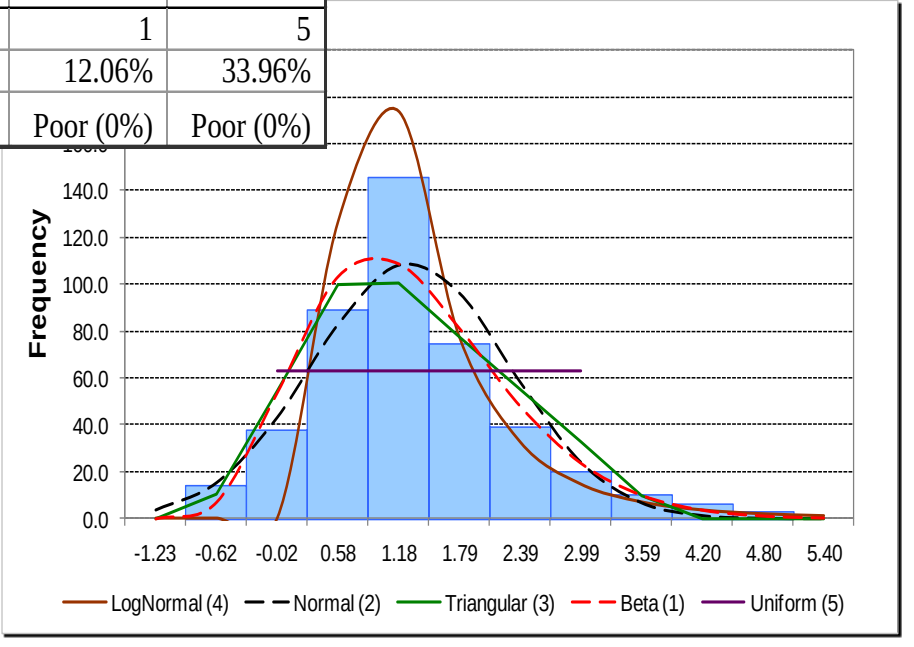
440 Normalized Percent Errors for USCM9 Subsystem-Level CERs (%error + 1)

Results derived by Distribution Finder for "adjusted % errors + 1":

	Sample	LN	Normal	Triangular	Beta	Uniform
Mean	1.0000	1.0801	1.0000	1.0000	1.0003	1.0000
StdDev	0.9746	0.8385	0.9565	0.9527	0.9674	0.9127
CV	0.9746	0.7763	0.9565	0.9527	0.9671	0.9127
Min	-1.2272			-1.0200	-1.3128	-0.5808
Mode		0.5323	1.0000	0.4655	0.6295	
Max	4.7993			3.5544	15.3924	2.5808
Alpha					4.7874	
Beta					29.7871	
Data Count	440	**% < 0 =**	14.79%	15.31%	14.45%	18.37%
Std Error of Estimate		0.3231	0.1888	0.2016	0.1206	0.3396
Rank		4	2	3	1	5
SEE / Fit Mean		29.92%	18.88%	20.16%	12.06%	33.96%
Chi^2 Fit test 22 Bins,	**Sig 0.05**	Poor (0%)	Poor (0%)	Poor (0%)	Poor (0%)	Poor (0%)

One is added to the normalized data to avoid centering on zero

1. Beta distribution fits the frequency histogram better than the other four distributions.
2. None of these five distributions pass the Chi-square test.



440 Normalized Percent Errors for USCM9 Subsystem-Level CERs (%error + 3.8231)

Results derived by Distribution Finder for "adjusted % errors + **3.8231**":

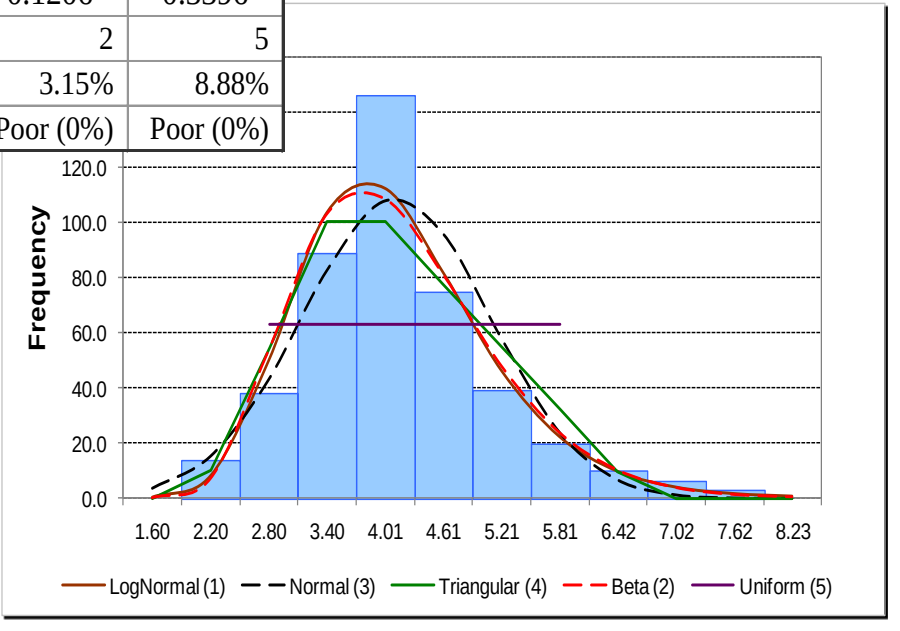
	Sample	LN	Normal	Triangular	Beta	Uniform
Mean	3.8231	3.8235	3.8231	3.8231	3.8234	3.8231
StdDev	0.9746	0.9709	0.9565	0.9527	0.9674	0.9127
CV	0.2549	0.2539	0.2502	0.2492	0.2530	0.2387
Min	1.5959			1.8031	1.5126	2.2423
Mode		3.4814	3.8231	3.2886	3.4520	
Max	7.6224			6.3775	18.2559	5.4039
Alpha					4.7803	
Beta					29.8555	
Data Count	440	% < 0 =	0.00%	None	None	None
Std Error of Estimate		0.1011	0.1888	0.2016	0.1206	0.3396
Rank		1	3	4	2	5
SEE / Fit Mean		2.64%	4.94%	5.27%	3.15%	8.88%
Chi^2 Fit test 22 Bins,	**Sig 0.05**	Poor (3%)	Poor (0%)	Poor (0%)	Poor (0%)	Poor (0%)

This example illustrates the **shifted** log-normal distribution is more useful than LN(u,σ^2,0).

Solver is used to find a location parameter when fitting a regular log-normal distribution (LN(u,σ^2,0)).

3.8231 is an **average** location parameter for these 8 subsystems. It is not a meaningful number.

1. **LN** distribution fits the frequency histogram better than the other four distributions, but none of them pass the Chi-square test**.**
2. **LN** distribution has a standard deviation of 0.25 in log space, which is smaller than the smallest SPE of all the eight subsystem CERs under investigation. The fitted results are doubtful.



Use Distribution Finder to model the error distribution for USCM9 Subsystem-Level CERs

No specific locations are considered in the analysis, as it is a generalized assessment

USCM9 Attitude Control System CER % Errors ($y_i/\hat{y}_i$)

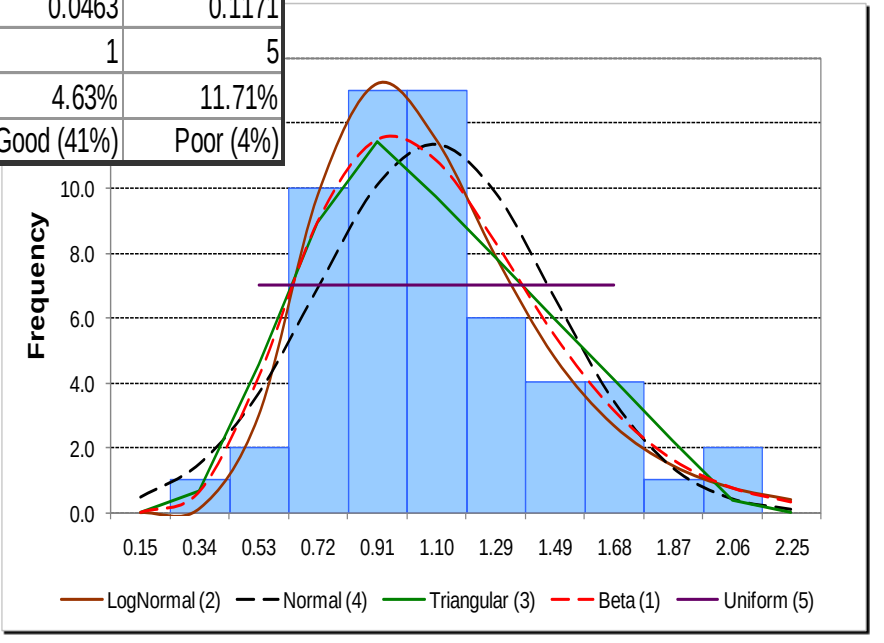
Results derived by Distribution Finder for the ratios of $y_i/\hat{y}_i$:

	Sample	LogNormal	Normal	Triangular	Beta	Uniform
Mean	1.0000	1.0039	1.0000	1.0001	1.0008	1.0000
StdDev	0.3776	0.3732	0.3722	0.3698	0.3761	0.3562
CV	0.3776	0.3718	0.3722	0.3697	0.3758	0.3562
Min	0.1490			0.2359	0.0684	0.3830
Mode		0.8268	1.0000	0.7637	0.8636	
Max	2.0583			2.0006	6.4772	1.6170
Alpha					5.1081	
Beta					29.9987	
Data Count	56	% < 0 =	0.36%	None	None	None
Standard Error of Estimate		0.0521	0.0696	0.0645	0.0463	0.1171
Rank		2	4	3	1	5
SEE / Fit Mean		5.19%	6.96%	6.45%	4.63%	11.71%
Chi^2 Fit test 10 Bins, Sig 0.05		Good (74%)	Good (17%)	Good (41%)	Good (41%)	Poor (4%)

a. Raw percent errors (i.e., $y_i/\hat{y}_i$) are analyzed by Distribution Finder. No correction factors are applied due to large sample size.

b. These raw % errors are not normalized, as they are from the same subsystem.

1. Both **Beta** and **LN** distributions fit the frequency histogram reasonably well.
2. Uniform distribution does not pass the Chi-square test (the other four pass the test).
3. **Beta** and **LN** distributions seem to be popular candidates to model the CER uncertainties.



USCM9 Propulsion CER % Errors ($y_i/\hat{y}_i$)

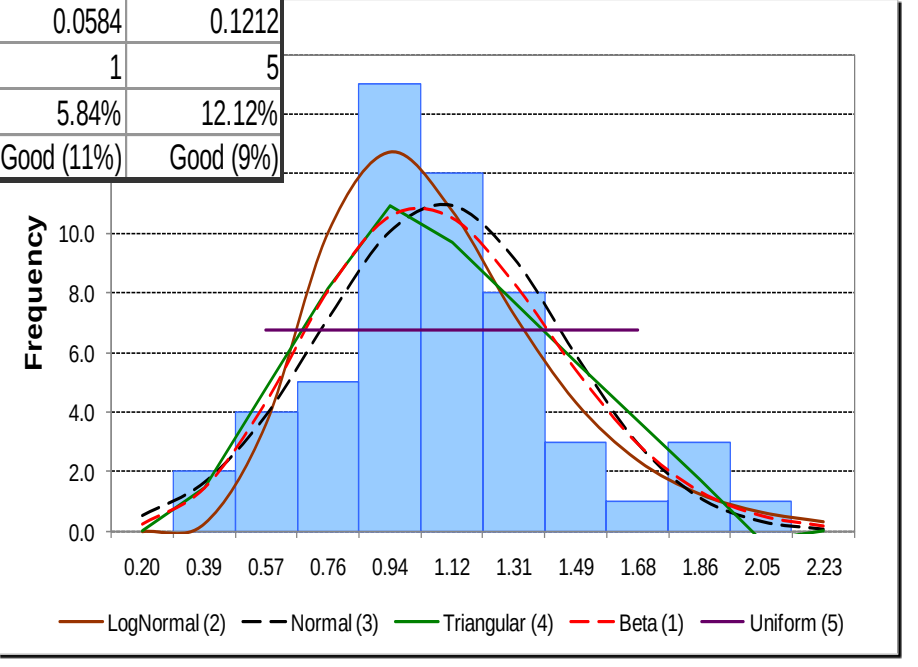
Results derived by Distribution Finder for the ratios of $y_i/\hat{y}_i$:

	Sample	LogNormal	Normal	Triangular	Beta	Uniform
Mean	1.0000	1.0038	1.0000	1.0000	1.0004	1.0000
StdDev	0.3620	0.3550	0.3570	0.3523	0.3578	0.3384
CV	0.3620	0.3536	0.3570	0.3523	0.3576	0.3384
Min	0.2047			0.2185	-0.3405	0.4139
Mode		0.8412	1.0000	0.8556	0.9343	
Max	2.0452			1.9261	4.7226	1.5861
Alpha					10.0616	
Beta					27.9286	
Data Count	54	% < 0 =	0.25%	None	0.01%	None
Standard Error of Estimate		0.0624	0.0657	0.0735	0.0584	0.1212
Rank		2	3	4	1	5
SEE / Fit Mean		6.22%	6.57%	7.35%	5.84%	12.12%
Chi^2 Fit test 10 Bins, Sig 0.05		Good (84%)	Good (28%)	Good (20%)	Good (11%)	Good (9%)

a. Raw percent errors (i.e., $y_i/\hat{y}_i$) are analyzed by Distribution Finder. No correction factors are applied.

b. These raw % errors are not normalized, as they are from the same subsystem.

1. Both **Beta** and **LN** distributions fit the frequency histogram reasonably well.
2. All five distributions pass the Chi^2 test.
3. **Beta** and **LN** distributions seem to be popular candidates to model the CER uncertainties.



USCM9 Electrical Power Subsystem CER % Errors ($y_i/\hat{y}_i$)

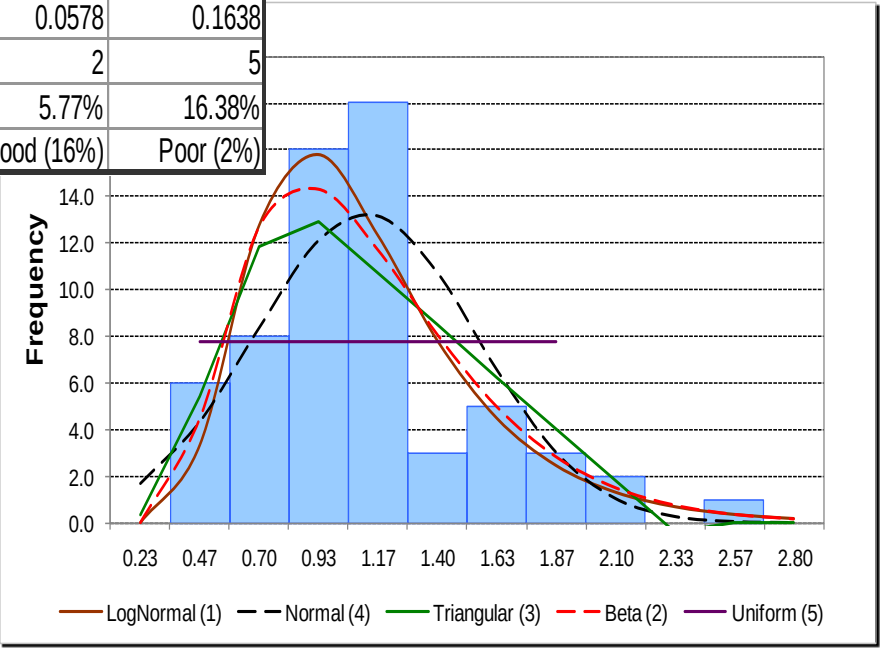
Results derived by Distribution Finder for the ratios of $y_i/\hat{y}_i$:

	Sample	LogNormal	Normal	Triangular	Beta	Uniform
Mean	1.0000	1.0037	1.0000	1.0001	1.0013	1.0000
StdDev	0.4438	0.4458	0.4308	0.4297	0.4427	0.4097
CV	0.4438	0.4441	0.4308	0.4296	0.4421	0.4097
Min	0.2315			0.1556	0.2236	0.2904
Mode		0.7662	1.0000	0.6654	0.7501	
Max	2.5675			2.1792	9.5042	1.7096
Alpha					2.7440	
Beta					30.0000	
Data Count	62	% < 0 =	1.01%	None	None	None
Standard Error of Estimate		0.0489	0.1111	0.1016	0.0578	0.1638
Rank		1	4	3	2	5
SEE / Fit Mean		4.87%	11.11%	10.16%	5.77%	16.38%
Chi^2 Fit test 10 Bins, Sig 0.05		Good (33%)	Good (17%)	Good (18%)	Good (16%)	Poor (2%)

a. Raw percent errors (i.e., $y_i/\hat{y}_i$) are analyzed by Distribution Finder. No correction factors are applied due to the large sample size.

b. These raw % errors are not normalized, as they are from the same subsystem.

1. Both **Beta** and **LN** distributions fit the frequency histogram reasonably well.
2. Uniform distribution fails the Chi^2 test, but the other four pass.
3. **Beta and LN** distributions seem to be popular candidates to model the CER uncertainties.



Conclusions

- **Sample size can be a concern when using a distribution fitting tool**
- **Suggest fitting (1) residuals for additive error models, (2) percent errors in the form of ratios (i.e., $y_i / \hat{y}_i$) for MUPE and ZMPE CERs, (3) residuals in log space for log-error models, and (4) ratios of actual to the mean ($y_i / \bar{y}$) for univariate analysis**
- **Consider three adjustment factors when using a distribution fitting tool for cost uncertainty analysis: DF, regression, and location factors**
  - Do not apply the DF factor when the sample size is fairly large (e.g., DF > 50) or when a Student's t or a Log-t distribution is used to model the CER errors
  - Define a shift factor (1) for MUPE/ZMPE CERs, so the CER errors are centered on one and (2) for univariate analysis, so the sample mean stays the same
- **Do not pool all the residuals (or percentage errors) from various CERs to analyze them together using a distribution finding tool**

Recommendations and Future Study

- **Enrich distribution gallery**
  - Besides commonly used distributions, consider including the following distributions: Student's t, Log-t, Weibull, Shifted Log-Normal, Gamma, Extreme Value distribution, User-Defined Cumulative Distribution Function (CDF), etc.
- **Examine whether we should adjust DF for additional constraints**
  - If constraint is specified for the unknown parameters, then one restriction is probably equivalent to a gain of one DF
  - Should the inequality constraints be adjusted? If yes, how do we adjust them?
- **Consider applying User-Defined CDF to model sample data with two or multiple modes**
- **Additional research for Beta and Log-Normal distributions: can the "world" be described by Beta and Log-Normal?**

References

1. Smith, A., “Build Your Own Distribution Finder,” 2010 ISPA/SCEA Joint Annual Conference, San Diego, CA, 8-11 June 2010.
2. Nguyen, P., B. Kwok, et al., “Unmanned Spacecraft Cost Model, Ninth Edition,” U. S. Air Force Space and Missile Systems Center (SMC/FMC), Los Angeles AFB, CA, August 2010.
3. Hu, S., “The Minimum-Unbiased-Percentage-Error (MUPE) Method in CER Development,” 3rd Joint Annual ISPA/SCEA International Conference, Vienna, VA, 12-15 June 2001.
4. Book, S. A. and N. Y. Lao, “Minimum-Percentage-Error Regression under Zero-Bias Constraints,” Proceedings of the 4th Annual U.S. Army Conference on Applied Statistics, 21-23 Oct 1998, U.S. Army Research Laboratory, Report No. ARL-SR-84, November 1999, pages 47-56.
5. Seber, G. A. F. and C. J. Wild, “Nonlinear Regression,” New York: John Wiley & Sons, 1989, pages 37, 46, 86-88.

Backup

Data Set

Observations	Cost	Weight
1	3,793.99	611.38
2	10,676.77	2,327.69
3	10,524.52	2,285.41
4	9,095.45	2,177.02
5	1,511.39	365.58
6	2,322.08	1,663.67
7	775.56	244.39
8	1,903.68	488.10
9	2,741.95	582.79
10	2,006.08	460.80
11	1,493.83	448.28
12	5,866.79	1,352.53
13	4,772.77	1,309.53
14	2,568.28	430.67
15	2,773.27	477.77
16	2,234.63	582.08
17	1,135.33	244.31
18	1,727.35	461.44
19	3,563.26	599.59
20	30,984.36	15,389.49
21	1,492.72	461.86
22	10,073.28	1,457.76
23	4,050.03	689.19
24	5,616.32	1,096.39
25	2,359.17	825.43
26	3,287.88	641.93
27	2,618.42	750.25
28	2,358.90	668.90
29	6,287.16	1,843.79
30	5,510.83	1,250.26
31	3,572.06	1,053.85
32	3,010.41	1,053.85
33	7,165.43	1,780.29
34	6,475.55	1,841.10
35	4,000.64	963.93
36	4,786.95	1,250.26
37	4,693.68	852.71
38	3,027.72	871.23
39	934.63	264.15
40	715.88	160.30
41	6,566.85	1,777.46
42	30,980.07	8,709.75
43	3,948.45	982.39
44	3,366.42	692.27
45	3,883.15	1,301.69
46	1,899.63	1,687.26
47	3,847.98	727.93
48	3,911.74	635.58
49	5,557.74	1,576.33